基于基函数展开的函数型数据聚类方法

陈晓锋，殷瑞飞
（厦门大学 经济学院 计划统计系，福建 厦门 361005）

摘 要：文章在一个一般性的框架下研究了利用基函数展开进行函数型数据聚类的问题。在这个框架之下，大量传统的聚类方法都可以直接应用到函数型数据分析。另外，我们将 Pearson 相似系数引入函数型数据聚类分析，解决了欧式距离无法刻画曲线之间形态差异的问题。

关键词：函数型数据；聚类分析；基函数
中图分类号：C812 文献标识码：A 文章编号：1002-6487(2009)19-0010-03

0 引言

在现实世界中，大量的数据具有函数性特征，例如人的生长曲线、导弹飞行的轨迹、气温的变化等。事实上，几乎所有时间序列数据都可以看成是关于时间的函数型数据，尽管我们只能得到有限数量的离散观测值。这种数据表面上看起来与面板数据极为相似，但实际上面板数据仅仅是函数型数据的一种类型。

Ramsay(1982)首先提出函数型数据的概念以及函数型数据分析(FDA)的研究思路和方法框架。FDA 的基本特征是把数据看成一个完整的函数对象，而不是个体观测值的简单排列。

近年来，许多传统的多元统计方法已经被扩展到了函数型数据分析领域。Ramsay 和 Silverman(1997)将回归分析、主成分分析、判别分析、典型相关分析和广义线性模型等方法引入 FDA。Escabias 等(2000)提出了函数型 Logit 回归方法。James(2002)将 Ramsay 和 Silverman 的广义线性模型加以深化。

将聚类方法应用于函数型数据的研究进展则相对缓慢。Gareth 和 Catherine(2003)使用三次样条基展开原函数，并把基展开系数看成服从联合正态分布的随机向量，建立了一种基于统计模型的函数型数据聚类方法。Abraham 等(2003)利用 B 样条基展开系数直接替代原函数，对奶酪发酵过程的函数数据进行 k-means 聚类。但是，由于 B 样条基函数是非正交基，这种做法将得到不一致的结论。Tarpey 和 Kinateder(2003) 提出了一种基于主点的函数型数据聚类方法。与 Abraham 等相比，Tarpey 和 Kinateder 的一个进步之处在于：他们指出了要想得到一致的结果，必须将原函数在相互正交的基函数上展开。但是，他们并没有给出一个严格的数学证明，而且也没有给出非正交基函数展开问题的解决办法。

本文拟在一个一般性的框架下研究利用基函数展开进行函数型数据聚类的问题。并将证明，只要将原始函数在标准正交基函数上展开，则原始函数之间的欧式距离与基展开系数向量之间的欧式距离是一致的；如果在非正交的基函数上展开，则只需对系数向量之间的距离进行一定修正，依然可以得到一致的结论。在这个一般的框架之下，大量传统的聚类方法都可以直接应用到函数型数据分析。另外，我们拟将 Pearson 相似系数引入函数型数据聚类分析，以解决欧式距离无法刻画曲线之间形态差异的问题。

1 函数数据的基函数展开

FDA 总是将一系列函数(曲线)作为研究对象，而我们实际上只能得到函数在有限时点上的取值。因此，FDA 的首要工作就是将离散的观测值转变成连续且光滑的函数形式。假设 $x^i(t)(i=1,2,\cdots,n)$是要研究的 n 个函数对象，而得到 $x^i(t)$的 T_i 个观测值 $y^i=(y^i_1,y^i_2,\cdots,y^i_{T_i})'$。由于有观测误差的存在，所以有下面的模型：

$$y^i_j=x^i(t^i_j)+\varepsilon^i(t^i_j)\ (i=1,2,\cdots,n;j=1,2,\cdots,T_i) \tag{1}$$

为了估计 $x^i(t)$，首先需要将其在一组基函数 $\Phi(t)=\{\phi_1(t),\phi_2(t),\cdots,\phi_K(t)\}'$ 上展开。即将 $x^i(t)$表示成基函数的线性组合：

$$x^i(t)=\sum_{k=1}^{K}c^i_k\phi_k(t) \tag{2}$$

矩阵形式为：

$$x^i(t)=c^{i\prime}\Phi(t),c^i=(c^i_1,c^i_2,\cdots,c^i_K)' \tag{3}$$

在每个时间点 t^i_j 上将式(2)代入式(1)，并应用最小二乘法则，就可以得到函数 $x^i(t)$的基展开系数向量 c^i：

$$c^i=\underset{c^i}{\operatorname{argmin}}\sum_{j=1}^{T_i}[y^i_j-\sum_{k=1}^{K}c^i_k\phi_k(t^i_j)]^2=(B^{i\prime}B^i)^{-1}B^{i\prime}y^i \tag{4}$$

其中，矩阵 $B^i=(\phi_k(t^i_j))_{T\times K}$ 中的元素是第 k 个基函数在时间点 t^i_j 上的取值。

在这里我们看到，每一个函数对象 $x^i(t)$都是利用各自的观测向量独立地进行估计它并不要求每个样品在相同的时点采样。同时，一旦给定基函数，则函数集合$\{x^1(t),x^2(t),\cdots,x^n(t)\}$的信息就被系数向量集合$\{c^1,c^2,\cdots,c^n\}$唯一地反映出来。

2 函数型数据聚类

2.1 基于基函数展开的函数型数据聚类

确定聚类对象间的距离度量指标是任何聚类分析的首要问题，对函数型数据的聚类也不例外。对于给定的两个函数 x(t)和 z(t)，衡量其距离的常用指标有差异的上确界、一致差异、欧式距离等。由于欧式距离具有优良的数学性质，因此成为衡量函数相似性的最常用指标（Abraham 等，2003；Tarpey 和 Kinateder，2003），其表达式如下：

$$D_{xz}=\int_0^T (x(t)-z(t))^2dt \tag{5}$$

如果直接依据式(5)计算两两曲线之间的欧式距离，则聚类过程将需要进行大量的数值积分，从而大大增加算法的时间复杂度①。为了简化运算，一个直观的想法就是用基函数展开系数向量的距离代替原函数之间的距离。为了证明二者之间的联系，将式(5)中的两条曲线用相同的 K 维基函数 Φ(t)展开，用 x 和 z 分别表示函数 x(t)和 z(t)的基函数展开系数向量，则有

$$D_{xz}=(x-z)'\int(\Phi(t)\Phi(t)')dt(x-z) \tag{6}$$

令 K 阶方阵 $W=\int(\Phi(t)\Phi(t)')dt$，则有

$$D_{xz}=(x-z)'W(x-z) \tag{7}$$

当我们选择的基函数是标准正交基时，矩阵 W 就退化成单位阵。显然，这时函数之间的距离就变成了系数向量之间的欧式距离。这样，原始函数之间的欧式距离与基展开系数向量之间的欧式距离是一致的。而当基函数非正交时，应被理解为系数向量之间以基函数的协差阵为权重的加权欧式距离。由于基函数的维数 K 一般不会很大，因此，矩阵 W 中的少量积分运算并不会带来运算速度上的麻烦。

于是，函数型数据的聚类分析就被转化成低维空间中系数向量的聚类分析，具体而言包括两个主要步骤：(1)将原函数利用某种基函数展开，得到一系列系数向量；(2)采用一定的聚类方法，依据加权或未加权的欧式距离，对系数向量进行聚类。

这样，我们就得到一个解决函数型数据聚类问题的一般框架。其一般性体现在两个方面：(1)原函数可以利用任意基函数展开，无论该基函数是正交的(如傅立叶基)，还是非正交的(如 B 样条基)；(2)任何基于欧式距离的传统聚类方法(如 k-means、系统聚类法等)都可以被应用到函数型数据分析中。

2.2 依据相似系数进行聚类

欧式距离的一个巨大缺陷是它仅仅衡量了曲线之间的位置差异，而没有捕捉到曲线之间的形态差别②。为此，我们将 Pearson 相似系数指标引入函数之间的相似性度量：

$$\rho_{xz}=\frac{\int_0^T[x(t)-\frac{1}{T}\int_0^T x(t)dt][z(t)-\frac{1}{T}\int_0^T z(t)dt]dt}{\sqrt{\int_0^T[x(t)-\frac{1}{T}\int_0^T x(t)dt]^2dt\int_0^T[z(t)-\frac{1}{T}\int_0^T z(t)dt]^2dt}} \tag{8}$$

可以看出，Pearson 相似系数的计算过程本身已经包含了对曲线的标准化过程，消除了曲线绝对水平高低的影响，从而突出了曲线的形态特征。

为了避免数值积分运算，将函数 x(t)和 z(t)用基函数展开，设其系数向量分别为 x 和 z。令 K 维向量 $u=\int_0^T\Phi(t)dt$，则有

$$\int_0^T\Phi(t)dt=x'\int_0^T\Phi(t)dt=x'u \tag{9}$$

$$\int_0^T[x(t)-\frac{1}{T}\int_0^T x(t)dt]^2dt=x'Wx-\frac{1}{T}x'uu'x$$

$$\int_0^T z(t)dt=z'u$$

$$\int_0^T[z(t)-\frac{1}{T}\int_0^T z(t)dt]^2dt=x'Wx-\frac{1}{T}z'uu'z$$

$$\int_0^T[x(t)-\frac{1}{T}\int_0^T z(t)dt][z(t)-\frac{1}{T}\int_0^T z(t)dt]dt=x'Wz-\frac{1}{T}z'uu'z \tag{10}$$

代入式(8)可以得到

$$\rho_{xz}=\frac{xWz-\frac{1}{T}z'uu'z}{\sqrt{(x'Wx-\frac{1}{T}x'uu'x)(z'Wz-\frac{1}{T}z'uu'z)}} \tag{11}$$

与矩阵 W 类似，向量 u 中的少量积分运算并不会带来运算速度上的麻烦；而且，对于大部分基函数，向量 u 具有解析解。

3 实证模拟

我们选用 2005 年 2 月 28 日至 2006 年 2 月 24 日中证 100 指数中具有连续交易数据的 82 家上市公司股票的日收盘价作为分析对象，把每支股票的价格经过对数化处理后，使用 18 维 B 样条基函数拟合曲线。图 1 是 82 支股票价格的拟合曲线及其一阶导数曲线。由于股价经过对数化处理，所以，一阶导数曲线实际上就是收益率曲线。图中大量曲线交杂在一起，因此很难直接从图中看出股票的类别信息。

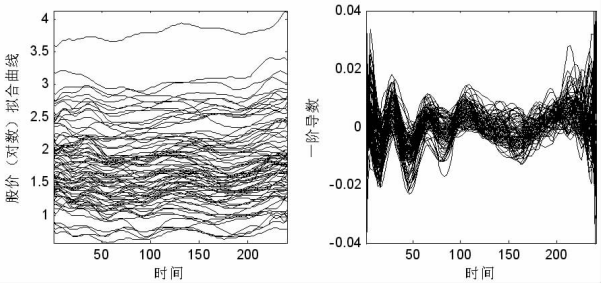

图 1 股价拟合曲线(左)与一阶导数曲线(右)

(注：共 82 支股票，241 个交易日，价格使用复权价，且经过对数化处理，数据来源于 Wind 数据库。)

①例如，若使用系统聚类法，则时间复杂度为 $O(n^2m)$，即使是使用算法速度较快的 k-means 法，时间复杂度也需 O(nktm)。其中，n 为样本容量，m 表示数值积分的分段数，k 表示 k-means 法的分类数目，t 表示 k-means 法的迭代次数。

②正是由于这个原因，Marron 和 Tsybakov(1995)、Heckman 和 Zamar(2000)分别构造了基于形态的相似性指标和基于秩相关的相似性指标来刻画曲线之间的形态差别。然而，这两种指标的计算都涉及到大量数值积分，因此运算速度成了它们的瓶颈。

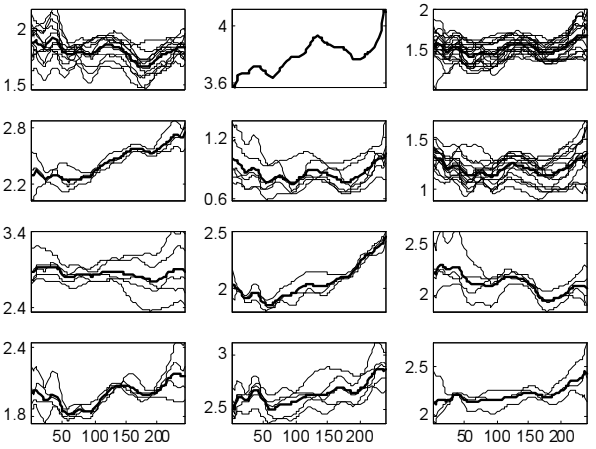

图 2 依据欧式距离进行聚类的结果

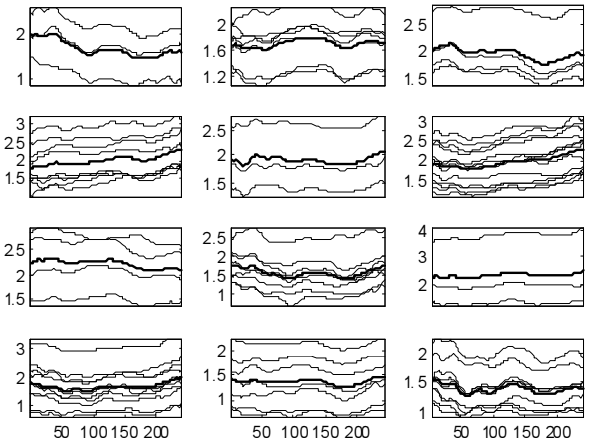

图 3 依据相似系数进行聚类的结果

首先采用欧式距离作为函数间相似性度量，我们以 k-means 为例来说明本文的聚类方法③。由上文可知，这可以通过对 82 个 18 维向量的 k-means 聚类来实现④，图 2 显示了以欧式距离作为相似性度量时将 82 支股票聚为 12 类的结果，图中的粗线条表示各类的均值函数。从图中可以看出，这里的聚类结果基本上是将绝对价格水平相近的股票聚在了一起(这一点可以从各个子图的纵坐标看出)，而并没有很好地捕捉到股价"走势"的相似性。

为了消除股价绝对水平的影响，使得聚类结果真正体现曲线的形态特征，我们以式(11)作为相似性指标，依然利用 k-means 聚类法，对 82 支股票走势重新进行聚类。其算法步骤与以欧式距离为依据的 k-means 聚类大同小异，差别仅在于步骤 2 中需要根据式(11)计算每个系数向量与各个类均值向量的相似系数，并按照相似系数最大原则进行重新分配。聚类结果如图 3。从图中可以看出，水平因素的影响已经被消除，聚类结果很好地捕捉了股价在形态上的相似性。

4 结束语

传统的聚类方法只能解决静态问题，函数型数据聚类分析则可以从动态的角度描述事物的类别，因此大大扩展了聚类分析的理论框架及其应用范围。

本文通过函数型数据的基函数展开，将函数型数据的聚类转化为低维空间中系数向量的聚类问题，并在此基础之上建立了一个解决函数型数据聚类问题的一般框架。在这个框架之下，原函数可以利用任何基函数展开，且任何基于欧式距离或相似系数的传统聚类方法都可以被应用到函数型数据分析中。并且，我们将 Pearson 相似系数引入函数型数据聚类分析，实证结果表明它可以较好地捕捉数据的相似特征。

参考文献：

[1]Abraham C., Cornillon P. A., Matzner-Lober E., Molinari N.Unsupervised curve clustering using B-splines [J].Scandinavian Journal of Statistics,2003,30.

[2]Escabias M., Aguilera A. M. and Valderrama M. J.Principal Component Estimation of Functional Logistic Regression: Discussion of two Different Approaches [J].Journal of Nonparametric Statistics,2004,16.

[3]Evans J. F., Arch S. H. Diversification and the Reduction of Dispersion: An Empirical Analysis[J].Journal of Finance,1968,23(5).

[4]Gareth M. J. Catherine A. S., Clustering for Sparsely Sampled Functional data[J].Journal of American Statistical Association,2003,98.

[5]Heckman N. E. Zamar R. H., Comparing the Shapes of Regression Functions[J].Biometrika,2000,87.

[6]James G. M., Generalized Linear Models with Functional Predictors[J].Journal of the Royal Statistical Society,2002,64(3).

[7]Johnson K. H. Shannon D. S., A Note on Diversification and the Reduction of Dispersion[J].Journal of Financial Economics,1974,4.

[8]Marron J. S. Tsybakov A. B., Visual Error Criteria for Qualitative Smoothing [J].Journal of American Statistical Association,1995,90.

[9]Ramsay J. O.When the Data Are Functions [J].Psychometrika, 1982,47(4).

[10]Ramsay J. O. Silverman B. W. ,Functional Data Analysis [M]. New York: Springer, Berlin Heidelberg,1997.

[11]Tarpey T. Kinateder K.K.J. Clustering Functional Data[J].Journal of Classification,2003,20.

(责任编辑/亦 民)

③正如前文所述，在这个一般性框架下，任何基于欧式距离的传统聚类方法都可以被应用到函数型数据分析。例如，若使用系统聚类法，则可以首先根据式(7)生成函数之间的距离矩阵，然后在此距离矩阵的基础上进行系统聚类。

④严格地讲，这里需要用到两个性质：

性质 1：类均值函数 $\bar{x}_h(t)$ 的基展开系数向量等于本类中函数基展开系数向量的均值 $\bar{x}_h$。证明如下：

$$\bar{x}_h(t)=\frac{1}{m_h}\sum_{j=1}^{m_h}x_{hj}(t)=\frac{1}{m_h}\sum_{j=1}^{m_h}x'_{hj}\Phi(t)=\frac{1}{m_h}\sum_{j=1}^{m_h}x'_{hj}\Phi(t)=\bar{x}'_h\Phi(t)$$

性质 2：各函数与类均值函数的欧式距离等于函数基展开系数向量与类均值向量的(加权)欧式距离。性质 2 可以根据性质 1 和式(6)很容易地得到证明。